To do list:

1. Methods er bade I nutid og datid. Ret det til det rigtige??
2. Provide graphs and plots and flow charts for methods section
3. Har jeg lidt discussion i mit methods? ”*It is good that I have data from different studies*”
4. Provide nice plots for results?
5. Read through and see if any ‘notes to section’ should be included in the sections, depending on space. This in OneNote!
6. Should I use “predicted audiofiles instead of participants” from notes in discussion?
7. Should I include ROC curve? <https://towardsdatascience.com/the-3-most-important-composite-classification-metrics-b1f2d886dc7b>
8. Remember to include link to github, for open science reasons
9. F1-scores and precision and recall are percentages
10. Write about “software used” somewhere in the methods section
11. Explain why random forest/cvsm/others is good in the paper.
12. Remember to evaluate accuracy not to chance but to majority group
13. Make sure Pipeline contains information on moving away from “accuracy”
14. Check if I need more in introduction on the specifics of my machine learning
15. Mention software VSCODE, Python, etc.
16. Mention all relevant places that: Even real diagnoses might be different (in discussion) – if thought of as continuum, perhaps only more schizophrenic people are diagnosed in Asia?? Or perhaps people diagnose differently across countries?
17. Read through papers in “papers downloaded” and add throughout where needed.
18. Skriv ML efter første gang det bliver nævnt, i stedet for machine learning
19. Should machine learning method (SVM linear kernel) be mentioned in detail?
20. What an ensemble model is and why we’re using it should be incorporated in the text
21. Nævn i methods et sted at vi arbejdede med predictions på voice-fil niveau og ikke participant!
22. Uddyb i methods at vi laver ”stacking” ensemble predictions. Så det bliver udspecificeret
23. Acknowledgements add or no? Perhaps also depends on Table of Contents (does it fit or not)
24. Can it really by utilized in the real world? (This model will believe that roughly 50% of population is schizo) – not going into depth with this.
25. Det originale studie bruger cross-validation
26. Overvej om forskelle mellem de 2 studier skal være til sidst i methods, eller når de skal bruges i discussion.
27. Check “Riccardo feedback - Voice atypicalities …” for his comments on discussion!
28. RICCARDOS NOTES:
29. “*you motivate your decisions, but don't say everybody should take the same ones. also you should make explicit that the procedure you are building is open for variations, so different things can be tried out.”*
30. “*Intro: define important concepts of ML*

*Methods: explain what you did in detali (eg we chose lasso regression)*

*How does it work to replicate the results of these paper when you increase the conservativeness of the procedure - and what do you learn from the problems, that future researcher can use when they do these kind of analyses*

*We need more research on how you developed principal ways of conducting analysis within this fields. Eg when you want to define your c parameter what approach should you take*

*-->and these problems are even bigger when you have a deep learning algorithm*

*parkinson's paper did this in a very good way:* [*https://asa.scitation.org/doi/full/10.1121/1.5100272*](https://asa.scitation.org/doi/full/10.1121/1.5100272)*“*

1. “*- make sure to have somebody else read the paper and point out where things are "indforstået", and generally give an eye on how fluent the language is. (hard to focus on that when you're still working on results and structure). Nothing major, but it does make a difference w censors.*”
2. “*- discuss somewhere how you would envision future developments of your pipeline: what are things to improve? How would one go replicating a different study? Etc.*”
3. “*- given the growing focus on symptoms and individual differences, would this pipeline/approach be scalable to that?”*
4. *“- feel free to send a more complete draft to Alberto Parola <a.parola@unito.it> for comments (cc'ing me and mentioning I said to do that). More complete in that he's not been part of the process, so he'll need a bit more to be able to follow.*”
5. *“Yes. Somewhere (but I might have missed it) put a diagram or short summary of all the points where your study differs from the original one.*”

# Abstract

Can machine learning (ML) applied to voice data be used as a tool to help diagnose people with schizophrenia? Numerous studies have shown high ML accuracy when classifying schizophrenia, but the ways in which they do so differ widely, as concluded in the latest meta study within the field. Little work has been done to replicate these previous ML methods on new data, and there is currently no consensus on which methods should be used.  
This study replicated two promising ML studies on new data, using an improved validation technique and an inclusion of sensitivity and specificity rates. Accuracy rates found through replication were dissimilar to the original studies, with study X\* and study Y\* having overall accuracy rates for classification at 60% and 67%. In other words, a drop of 6 and 3 percentage points for the two studies, respectively. Through discussion, this study has found that the difference in scores in the replication points toward low ecological validity and robustness. The rest of the literature was also discussed, and I found that the widely heterogeneous results within the field indicate similar trends.  
As a consequence, this study has attempted to establish a ML pipeline less prone to the pitfalls of ML, with the intention of establishing a general procedure for future research. Finally, this paper advocates for a more open and cumulative scientific community.

How to write abstract:  
<https://blackboard.au.dk/bbcswebdav/pid-2793891-dt-content-rid-9152972_1/courses/BB-Cou-Hold-36086/L1%20-%20Getting%20started.pdf>  
p. 14 - p. 18

# 1. Introduction

## 1.1 Schizophrenia and biomarkers

### 1.1.1 Schizophrenia

**Actual paper:**

**Notes for section:**

Biomarkers - why voice good? -> shows part of social impairment.

### 1.1.2 Biomarkers and voice atypicalities

**Actual paper:**  
  
**Notes for section:**

Biomarkers - why voice good? -> shows part of social impairment.

1. Schizophrenia and voice in general
   1. 3 methods of studying
2. Two meta-studies -> the lit. is a mess

1.

History of the project:

Schizophrenia has certain distinctive features vocally. (Alogia, blunt affect, "poverty of speech", "latency of speech" etc.). This has been known since forever (Bleuler, 1911; Kraepelin, 1919).

 Voice atypicalities in SZ’s have always been known (Bleuler, 1911; Kraepelin, 1919).  
Schizophrenia has certain distinctive features vocally. Qualitatively the atypicalities have been described using numerous different terms (Alogia, blunt affect, "poverty of speech", "latency of speech", increased pauses, distinctive tone, intensity of voice etc.).

Voice atypicalities have been studied using 3 methods. Qualitative perceptual ratings, quantitative acoustic analysis and ML investigations.

Qualitative perceptual ratings have found robust differences between SZ and TD.

Quantitative acoustic analyses have found fewer robust differences, with varying effect sizes and sometimes direction.

Quantitative acoustic analyses have identified acoustic features on the basis of automated processes, leaving the assessment of the acoustic features more reliable. Using automation, the features of a set of voice data will identical over multiple feature detections, given the same feature detection hard- and software.  
Here, fewer robust differences were found with varying effect sizes and direction, depending on the features investigated (Cohen et al., 2014; https://www.biorxiv.org/content/10.1101/583815v4.full.pdf).

This is also in next section?:

Multivariate ML investigations have found promising results. Focus on minimizing out-of-sample-error instead of within sample-error as when using more traditional analyses, makes the applicability of the method more practically generalizable. It also allows for analyzing multiple features in conjunction. High correlation between almost all features (3.3, correlation <https://www.biorxiv.org/content/10.1101/583815v4.full.pdf> ).  
It does, however, not allow for transparency as to wherein the acoustic differences between SZ and HC lie.

2.

We don't know which features proves to have differences between SZ and TD

The litt. is a mess - results in different directions.

There's already a metastudy on Schizophrenia; which found atypicalities on different voice/speaking parameters - with varying effect sizes.

Large heterogeneity between studies.

More demanding tasks meant larger effect sizes.

## 1.2 Machine learning for detection of acoustic patterns

### 1.2.1 Prospects of machine learning in classifying schizophrenia

**Actual paper:**  
  
**Notes for section:**

1. Metatext and motivation for going into depth with machine learning
2. Allows for finding features (feature selection)
3. Allows for analyzing multiple features in conjunction
4. Promising findings (high accuracy in many studies)
5. Less interpretability but more practical applications (cheap)

Will not go into other ML things (gesticulation or others – beyond the scope of this paper)

1.

Although:  
Qualitative perceptual ratings have found relatively robust differences in voice between SZ and TD. Relying on raters to assess perceptual differences has some limitations. A feature such as “latency of speech” is interpretable and is partly going to be rated on the basis of human intuition – this requires comprehensive training for the rater. Moreover, the complex interplay between multiple acoustic features is hardly very accessible, even given proper and rigorous training.  
  
Therefore:

ML doesn’t have this problem.

2.

Feature selection; ridge, lasso, elasticnet

3.

Multivariate ML investigations have found promising results. Focus on minimizing out-of-sample-error instead of within sample-error as when using more traditional analyses, makes the applicability of the method more practically generalizable. It also allows for analyzing multiple features in conjunction. High correlation between almost all features (3.3, correlation <https://www.biorxiv.org/content/10.1101/583815v4.full.pdf> ).

4.

Multivariate ML investigations have found promising results. Focus on minimizing out-of-sample-error instead of within sample-error as when using more traditional analyses, makes the applicability of the method more practically generalizable. It also allows for analyzing multiple features in conjunction. High correlation between almost all features (3.3, correlation <https://www.biorxiv.org/content/10.1101/583815v4.full.pdf> ).

5.

Clinical application -> given schizophrenia, and given samtaleterapi or drugs, see how they're doing along the way by them talking every week on their phone.  
"*In addition, voice analysis may potentially allow to assess the response to psychosocial or pharmacological treatment over longer periods using objective and quantitative indices, and enhance the capability of clinicians to capture the complex relationship between emotion regulation, expressive behavior, social perception and cognitive and clinical features of the disorder (e.g. Ben-Zeev et al., 2017; Dahlgren et al., 2018; Tahir et al., 2019)*" (Parola, Fusaroli et. al 2019)Va [(Bush et al., 1998)](https://www.zotero.org/google-docs/?KFKj12).  
  
It does, however, not allow for transparency as to wherein the acoustic differences between SZ and HC lie.

ML can perhaps help with showing:  
a) Severity of schizophrenic symptoms  
b) Diagnosis, schizophrenia

Practical applications:   
Assisting tool for assessing diagnosis (Parola, Fusaroli et. al 2019)

On prediction of severity of clinical features from acoustic measures:  
(Püschel et al., 1998)

6 (extra):

Applicability of Bachelors project:

Meta-science, open science.

Assisting tool for assessing diagnosis (Parola, Fusaroli et. al 2019)

Clinical application -> given schizophrenia, and given samtaleterapi or drugs, see how they're doing along the way by them talking every week on their phone.

"*In addition, voice analysis may potentially allow to assess the response to psychosocial or pharmacological treatment over longer periods using objective and quantitative indices, and enhance the capability of clinicians to capture the complex relationship between emotion regulation, expressive behavior, social perception and cognitive and clinical features of the disorder (e.g. Ben-Zeev et al., 2017; Dahlgren et al., 2018; Tahir et al., 2019)*" (Parola, Fusaroli et. al 2019)

Companies interested in this (Lasse Hansen), Switzerland Internship on this in depression

### 1.2.2 Current limitations in the literature

**Actual:**

**Notes:**

* Overfitting
* Differences in methods, method quality and levels of transparency
  + Effect sizes of acoustic features is partially determined by task (difficulty)
* Lack of replications
  + Promising results
  + No validation between datasets
  + Language differences
  + No performance robustness measures

## 1.3 Alleviating current limitations

### 1.3.1 Through replications

**Actual paper:**  
  
**Notes for section:**

s

### 1.3.2 Through proper ML implementation

**Actual paper:**  
  
**Notes for section:**

1. Meta text to have a rød tråd
2. Establish the need for a change in research (from above section)
3. Idea of a pipeline
4. Explain how a pipeline might alleviate the problem. And cause the change that is needed

Set up a standard pipeline.  
(Very) general introduction to pipeline in introduction. Just the conceptual structure.  
  
In intro:

“explain the most basic and most important steps”

BUT … (not feature selection on full dataset) etc.

Mention confusion matrix

In methods:  
Take the individual conceptual steps from the intro, and explain in specific detail how we tailored this ML “to the conceptual steps”.

The proposed pipeline follows a relatively simple overall structure, which can be thought of as 8 steps. It is important to note, that the course of action within each step to a certain extent depends upon the specific data and classification problem. Moreover, it is important to note that documenting extensively throughout is important for increasing transparency. High transparency a) allows for future replications and b) increases the applicability of the work, since knowledge will be available about exactly under which conditions a given machine learning approach performs a certain way. The steps of the pipeline are as follows:   
**1) Data acquisition.** Acquiring the voice data for fitting a machine learning algorithm – this may be data that has been acquired either through recording, or from using already recorded data. **2) Data preprocessing.** Consists of cleaning the data; cutting away irrelevant speech (e.g. from interlocutors), noise removal and extracting features. It might also include applying data augmentation (e.g. applying noise, reverb etc.). **3) Data partitioning.** Partitioning the data into a training and a holdout set for testing. Typically, a split of roughly 80/20 is used. Generally, a larger holdout set means more precise knowledge of the performance of a given model. A larger training set on the other hand generally means better and more robust predictions. **4) Feature scaling and selection.** Feature scaling is necessary for all machine learning algorithms that are distance-based. In these cases, scale the train and holdout set separately. Use information from the training set (e.g. min. and max. values in min./max. normalization) to do the scaling on both sets. This ensures no leakage of information from the training set to the test set, while still providing a common scale without losing information or distorting the differences in the range of values. Feature selection entails selecting only the relevant machine learning parameters, that improves classification. Many different methods can be applied in order to obtain a relevant feature set. It may also be skipped if theory or some motivation dictates it. However, feature selection tends to generally improve the robustness of predictions, shorten training times, avoid the curse of dimensionality and it increases interpretability. **5), 6), 7) Model training, parameter tuning, model testing.** These three steps are intertwined, and the cycle may be repeated. Divide the training set up into training and test and then train and test the model. After seeing how it performs on the test set, you may tune the parameters and repeat the process until the predictions are optimal. This can be done advantageously by the implementation of cross validation. **8) Validation on hold-out set.** The model then predicts the holdout set and performance is evaluated using relevant metrics. To allow for further insights into the performance of a given model, evaluation metrics ought also to be calculated for subgroups of participants (e.g. men/women or nationalities).

An overview of the pipeline can be seen in figure 1.

Fig 1.

*Pipeline with workflow that is well suited for classification machine learning on voice.*

**Notes for section:**

* S

s

### 1.3.3 Thesis statement / purpose of paper

**Actual paper:**  
  
**Notes for section:**

Noone should just mindlessly replicate, when replicating!! Important to take all the necessary steps.

Overall goal of thesis:  
How does it work to reproduce?  
Increase conservative  
What are the important lessons we’re learning while we’re replicating

**Riccardos words on overall goal of thesis:**

How does it work to replicate the results of these paper when you increase the conservativeness of the procedure - and what do you learn from the problems, that future researcher can use when they do these kind of analysis

**Thesis statement idea 1 (Maries):**

This thesis aims to investigate the capabilities of existing machine-learning classifying individuals with ASD from acoustic features. We will review previous literature, extract strong voice-features and machine-learning models, and validate models on new data. We predict that support vector machine will achieve higher accuracy but will have less x and that naive bayes will x. Additionally, we predict that validation methods x,y,z will make results stronger in specific case/weaker generalization. By this, we will attempt to establish a procedure for machine-learning studies that achieve the most robust and ecologically valid measures.

**Thesis statement idea 2:**

This thesis aims to replicate two promising findings of machine learning classification of schizophrenia, using voice data. Since the literature on the area has very heterogeneous findings, I expect worse performance given the new data that I will test on. Given the inrobustness and low ecological validity of ML attempts, I will attempt to establish a ML pipeline less prone to the pitfalls of ML, with the intention of establishing a general procedure for future research.

Tour de Bachelors:

[https://docs.google.com/document/d/1qc3tDtAg6sc2-zfnxaxqKl\_WydK3AAgaDLb1V7QjTTU/edit?fbclid=IwAR1JB53UmJcDEI8GnXEEvA4PcuWXvVeX\_ZN43VEamHHxMWsHYdAR\_Wo3vKY#](https://docs.google.com/document/d/1qc3tDtAg6sc2-zfnxaxqKl_WydK3AAgaDLb1V7QjTTU/edit?fbclid=IwAR1JB53UmJcDEI8GnXEEvA4PcuWXvVeX_ZN43VEamHHxMWsHYdAR_Wo3vKY)

Pipeline:

[https://docs.google.com/document/d/1fbfpR5ZQiVTZYChzWut06fkXA9CQziMMJNtFOiF6Giw/edit?fbclid=IwAR3MU3OTehQ\_nVEk0nB8PihR\_0clhxrIbOPE5y\_v6X9PQoqRfFKRbZDiF7o#](https://docs.google.com/document/d/1fbfpR5ZQiVTZYChzWut06fkXA9CQziMMJNtFOiF6Giw/edit?fbclid=IwAR3MU3OTehQ_nVEk0nB8PihR_0clhxrIbOPE5y_v6X9PQoqRfFKRbZDiF7o)

# 4. Discussion

## 4.1 Results and replication comparison

### 4.1.1 Performance

**Actual:**

**Notes:**

### 4.1.1 Performance comparison to original study

1. Performance of models on test
   1. Evaluation metrics (F1, accuracy, precision + recall,)
      1. F1-score for model overall
      2. F1-scores for patients and controls respectively
      3. Accuracy (not very telling)
      4. Precision + Recall
   2. Between sexes
      1. Well balanced in replication
      2. No information in original paper
         1. Ought to be included
   3. Wrap up about results
      1. Similar results, but slightly better predictions in original
2. Why are there performance differences?
   1. Methods (as will be discussed in next section)

Maybe include:

1. Performance of models on train
   1. High performance
   2. Low generalizability due to overfitting
      1. Mention bad study that overfits

* These things in terms of what is being predicted (train/test/holdout) - better predictions when overfitting – obviously. To underline how much more shit the predictions will get when you’re rigorous. With increased conservatism, how good are the results really?
* Performance between groups
  + Sexes:
    - This: Balanced test-set (but perhaps a bit worse training)
    - Original: No information. But balanced dataset.
  + Nationalities
    - This: None
    - Original: No information. But unbalanced dataset
* Where do the differences come from?
  + Data, Feature selection, Methods
  + See next sections

#### 4.1.2.1 Data

* Language/nationality
  + Biased because of difference in labeling
    - This: Danish diagnostics
    - Original: Chinese, Malay, Indian diagnostics
  + Biased because of difference in language
    - This: Danish
    - Original: 3 Countries, with different languages
* Task
  + This: mid-level difficulty; description of triangles. No social component
  + Original: high-level difficulty; interview. Social component
* Data quantity
  + This: More participants with shorter recordings
  + Original: Fewer participants with longer recordings
* Sound quality
  + This: Difference in recording equipment
  + Original: Maybe?
* What contributed to the differences in performance? (If any)
  + Possibly all. Likely not sound quality to a large extent
* The difference in the participants native country meant that not only did the language spoken in the recordings differs, but also that they were not speaking their own native language. Moreover, the pool of schizophrenic participants was likely to vary between the original and this replication. This is because both diagnostic tools and psychologist and psychiatrist training are heterogeneous between countries to some extent. \* CITE \*. \*PASSER DET??\*

#### 4.1.2.2 Partitioning

* There is none in the original. They use the same data for both training and testing.
* Training set
  + This: Mostly balanced on gender. Mostly balanced on diagnosis
  + Original: Balanced on gender. Very unbalanced on diagnosis
* Holdout set
  + This: Very balanced on gender. Very balanced on diagnosis
  + Original: Balanced on gender. Very unbalanced on diagnosis

#### 4.1.2.3 Feature scaling

* This: Feature selection
  + This: LASSO - 5-fold
  + Original: PCA
  + Hard to replicate, given the sparse information on how PCA was used
    - Their feature selection method hard to follow
    - Could have been understood in two different ways
  + Specific feature selection method shouldn’t have a large impact on performance

#### 4.1.2.4 Feature selection

* Type of feature selection
  + This: LASSO - 5-fold
  + Original: PCA
  + Hard to replicate, given the sparse information on how PCA was used
    - Their feature selection method hard to follow
    - Could have been understood in two different ways
  + Specific feature selection method shouldn’t have a large impact on performance

Explanation of how it could be understood:

“*the features of the training set were ranked using one of the following techniques: F-score (ANOVA), χ 2 , Mutual Information, Pearson correlation, Principal Components, linear SVM, Decision Trees, and Random Forests. Subsequently, the optimal number of features were selected according to the previous ranking methods*”  
PCA used to rank? Most common method is that PCA is used for defining new features, namely PC1 + PC2 + ... +PCn, until some desired threshold of accumulated variance is met.

There’s also the possibility that it truly was used to rank, e.g. by looking at the features with least shared variance in the different principal components to avoid covarying features, but also here it is not possible to replicate 1-1. The method is still not specified

Shouldn’t really matter:

Regardless, of method used by Chakraborty et al, the method used here is good. And if the method using speech for classification truly is robust, then either would work. If these results truly are reliable and reliable, they shouldn’t be dependent on PCA/LASSO / whatever

Understanding PCA notes:

Link of idea of PCA for feature selection. (starts at 3:50). It shows that there are different methods (example with gain, here)

<https://www.youtube.com/watch?v=YEDOSOd44bU&list=PLBv09BD7ez_5_yapAg86Od6JeeypkS4YM&index=2&frags=wn&ab_channel=VictorLavrenko>

Link for example of PCA for feature selection (creating new features):

<https://www.quora.com/How-do-you-use-PCA-for-feature-selection>

#### 4.1.2.5 Machine learning algorithm

* Predicting (single participants, or same participants multiple times)
  + This: Predicting .wav files (several for each participant)
  + Original: Predicting participants
  + Should not have large impact on performance
* Ensemble modeling vs. Single machine learning algorithm
  + Stacking ensemble modeling
    - Better (if models are diverse, and generally good)
    - Only very slightly better
  + Single machine learning algorithm
    - Slightly worse

Should not have large impact on performance

Creating an ensemble model as opposed to using a single algorithm has the advantage of (possibly \* ) being more robust and reliable in its out-of-sample predictions (as mentioned in 4.1.1). Similarly, having the final testing set being

### 4.1.4 Methods ()

**Actual:**

**Notes:**

1. Predicting (single participants, or same participants multiple times)
   1. This: Predicting .wav files (several for each participant)
   2. Original: Predicting participants
   3. Does this matter?
2. Ensemble modeling vs. Single machine learning algorithm
   1. Stacking ensemble modeling
      1. Better (if models are diverse, and generally good)
      2. Only very slightly better
   2. Single machine learning algorithm
      1. Slightly worse

MAYBE INCLUDE:

* Specifics on ensemble modeling
  + Diversity/data trade-off in ensemble modeling

FOR BELOW I DO THE OPPOSITE NOW!!!!

Diversity/data trade-off in ensemble modeling:

I use all training data in each of the ensemble-sub-models. As opposed to excluding the test sets, that were also excluded for feature selection.

Could this be an issue?  
Yes; groups of diverse problem solvers (in general) outperform, the best (also often similar) models. At least when the diverse problem solvers and the better, more similar models have roughly the same amount of data. (Hong & Page, 2004)

Why did I choose to do it anyways?  
The “diversity” the opposing idea would bring, is not due to difference in neither type of models or any other diversity parameter. The opposing choice would only give diversity from differences in training data.

The increase in diversity would in this study, be on the cost of less training data. And less training data means worse predictions in general.

Does it really matter?

The trade-off between more/less training data and more/less diversity is unlikely to have had much of an impact. E.g. Less than 2 percent increase in acc. when having 10 agents (and we only have 5, which would probably mean even less of an impact) (Hong & Page, 2004).

But the potentially very small positive effect a more diverse set of decision-agents, might very well be negated by the fact that all of the 5 diverse models would be worse, due to their more limited data. In other words; no – it isn’t likely to have had a large effect. But it would have been interesting to do both.

## 4.2 Pipeline

### 4.2.1 (Narrow) How did an implementation of pipeline in this replication work out?

**Actual:**

**Notes:**

1. Meta
2. Data acquisition
   1. What did we do and why? Pros + cons + alternatives?
   2. Differences to original study? (if relevant)
3. Preprocessing
   1. What did we do and why? Pros + cons + alternatives?
   2. Differences to original study? (if relevant)
4. Data partitioning
   1. What did we do and why? Pros + cons + alternatives?
   2. Differences to original study? (if relevant)
5. Feature scaling
   1. What did we do and why? Pros + cons + alternatives?
   2. Differences to original study? (if relevant)
6. Feature selection
   1. What did we do and why? Pros + cons + alternatives?
   2. Differences to original study? (if relevant)
7. Model tuning (training, tuning and testing cycle)
   1. What did we do and why? Pros + cons + alternatives?
   2. Differences to original study? (if relevant)
8. Validation (and evaluation)
   1. What did we do and why? Pros + cons + alternatives?
   2. Differences to original study? (if relevant)
9. Reflection / evaluation + proper documentation
   1. Did we do this and why?

Differences to original study? (if relevant)

Given that this replication did not process the same data, nor used the same techniques for neither partitioning, feature scaling, feature selection, or for machine learning model, it is not surprising that the results differ (see table x \* for short summary).

The data acquisition step varied greatly as there were dissimilarities in the participant pool, the task and in both the length and number of recordings.  
It can be hypothesized that conditions such as alogia or the flat effect sometimes found in patients that are thought to elicit some of the acoustic atypicalities might manifest itself differently across languages. The fact that this replication had participants speak Danish as opposed to English might impact the ML algorithms ability to detect patterns for classification. Moreover, none of the participants spoke their first language in the original study given their Malay, Indian or Chinese origin. As of yet, research points towards some general differences in acoustic patterns in schizophrenia patients related to symptoms such as alogia and flat the effect\* Cite \*. However, from the knowledge of this researcher, very little research sheds light on the potential modulation that language or language nativeness might induce. Moreover, the pool of schizophrenic participants might also vary between the original and this replication as people diagnosed with schizophrenia elicit slightly different symptoms (*Lundbeck Institute Campus*, 2016; Sartorius et al., 1986).

The number of recordings was significantly higher in this replication given the large number of participants and the fact that each participant went through 8-10 trials with separate recordings. This meant that the feature extraction process produced more feature vectors (1 per recording) in this replication. In machine learning, each feature vector represents a data point and thus the classification algorithm simply had more datapoints to learn from. The recordings were however, substantially longer in the Chakraborty et al. study which meant that the feature vectors for each data point were more accurate and less prone to random variation \* cite \*.

As using SVM as an algorithm requires scaled parameters/features, this study employed a min-max normalization. The scaling of both the training and holdout set used the minimum and maximum values from the training set to ensure no information could flow from the training to the holdout set (Myrianthous, 2020). As no information was provided in the original paper, it is unclear whether their acoustic features were scaled within each step of the cross-validation, ensuring to scale the test set using only information from the training set, or if they scaled prior to the cross-validation process. The latter could result in a small amount of overfitting. Performance would be slightly better, but it would reflect out-of-sample performance as accurately. The reason for this would be that the classification algorithm could have learned from the testing data before seeing it for the validation (Géron, 2019).

LASSO regularization was utilized for feature selection in this study. Contrastingly, Chakraborty et al. utilized Principal Component Analysis (PCA). PCA reduces the dimensionality (number of features) of each data point (each recording), by generating a smaller number of new ‘principal components (dimensions) while preserving as much as the data’s variation as possible (Abdi & Williams, 2010). The latter feature selection technique diminishes the interpretability of the model, given that the original acoustic features are convoluted in the new principal components. Choosing one specific feature selection technique over another should in theory not have a large impact on performance in classification. Much theory supports the choice being arbitrary, but in practice it sometimes is not (Oreski et al., 2017). However, both PCA and LASSO have been found as some of the best feature selection techniques, with great improvements of classification algorithms (Sun et al., 2019). It is therefore unlikely that all the variation in performance between the two studies can be attributed solely to feature selection technique. If the method for using the acoustic features from ‘emobase’ for classification truly is robust and reliable, then either should work.

1. Replication
   1. Possible
   2. Hard (Methods explained in condensed manner)
2. Comparison
   1. Possible
   2. Hard (More information on sexes and nationalities needed)
3. Getting similar results
   1. Differences in performance – where does it come from?
      1. Biased labels
      2. Difference in language
      3. Task differences
      4. Difference in algorithms
      5. Arbitrary choices for tuning
      6. A mixture (which mixture?) of all the above
   2. Some things might balance each other’s out, some might not

### 4.2.2 Problems established from conservative replication

**Actual:**

**Notes:**

1. Curious that other studies have found much(!) higher accuracies
   1. Study 1
   2. Study 2
   3. Overfitting?
      1. My predictions on training 90% accuracy
      2. Scaling
2. Hard to know where differences in performance come from
   1. (All the differences on task, data, language, labeling etc.)
   2. Solution: More documentation on this and more reproductions
3. Bad documentation is insufficient for facilitating replication
   1. From practical experience
4. It is up to individual researchers and their experience to produce original studies and replications alike (not good)
   1. Arbitrary choices and handycrafts
      1. Tuning (C-parameters)
      2. Model type
   2. From practical experience – not possible to find established pipeline and solutions

To underline how much more shit the predictions will get when you’re rigorous  
With increased conservatism, how good are the results really?

Results real good if overfit

To underline how much more shit the predictions will get when you’re rigorous  
With increased conservatism, how good are the results really?

**Discussion:**

Many choices have to do with handycrafts and arbitrary choices (tuning)

We haven’t gotten enough research on principles of how to do this (Right now up to individual experience of researchers)

Present the issue that this hasn’t been fixed by our paper either. Even with relatively simple algorithms. Deep learning would mean that this is even worse.

So is language/way of speaking

* Hard to know whether differences in performance are due to:
  + biased labels
  + difference in language
  + task differences
  + difference in machine learning algorithm
  + arbitrary choices for tuning
  + a mixture of all (could balance each other out, if in opposite directions (some make it harder for the Danish corpus, some make it easier)

## 4.3 Further research

### 4.3.1 Need for a widely applicable, conservative, transparent pipeline.

1. Meta (widely applicable, conservative, transparent pipeline) would help by:
   1. Avoiding in overfitting (as mentioned previously)
   2. Making it easier to compare results (as mentioned previously)
      1. Within or across sexes and nationalities,
   3. Making it easier to replicate
   4. Enabling research to know locate the origin of differences in results (as mentioned previously)
      1. Biased labels
      2. Difference in language
      3. Task differences
      4. Difference in algorithms
      5. Arbitrary choices for tuning
      6. A mixture (which mixture?) of all the above
   5. Shedding light on arbitrary choices – either by:
      1. Providing information on it in the papers
      2. Providing a method for making these choices
2. In general: More replications and a generally more open-science based approach

This study was not enough.

**Discussion:**

Many choices have to do with handycrafts and arbitrary choices (tuning)

We haven’t gotten enough research on principles of how to do this (Right now up to individual experience of researchers)

Present the issue that this hasn’t been fixed by our paper either. Even with relatively simple algorithms. Deep learning would mean that this is even worse.

1. Establish a widely accepted pipeline. It should be:
   1. Rigorous
   2. Conservative
   3. Transparent
2. It allows for
   1. Replications
   2. Transparency

To test generalizability and robustness

This allows for more replications

# 5. Conclusion

**Actual paper:**

**Notes:**

# 6. Acknowledgements

s

# 7. References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1–16. https://doi.org/10.1016/S0885-2014(00)00014-9

Alberto, P., Arndis, S., Vibeke, B., & Riccardo, F. (2019). *Voice Patterns in Schizophrenia: A systematic Review and Bayesian Meta-Analysis* [Preprint]. Bioinformatics. https://doi.org/10.1101/583815

Beck, K. I., Simonsen, A., Wang, H., Yang, L., Zhou, Y., & Bliksted, V. (2020). Cross-cultural comparison of theory of mind deficits in patients with schizophrenia from China and Denmark: Different aspects of ToM show different results. *Nordic Journal of Psychiatry*, 1–8.

Bliksted, V., Fagerlund, B., Weed, E., Frith, C., & Videbech, P. (2014). Social cognition and neurocognitive deficits in first-episode schizophrenia. *Schizophrenia Research*, *153*(1), 9–17. https://doi.org/10.1016/j.schres.2014.01.010

Bliksted, V., Frith, C., Videbech, P., Fagerlund, B., Emborg, C., Simonsen, A., Roepstorff, A., & Campbell-Meiklejohn, D. (2019). Hyper-and hypomentalizing in patients with first-episode schizophrenia: FMRI and behavioral studies. *Schizophrenia Bulletin*, *45*(2), 377–385.

Chakraborty, D., Yang, Z., Tahir, Y., Maszczyk, T., Dauwels, J., Thalmann, N., Zheng, J., Maniam, Y., Amirah, N., & Tan, B. L. (2018). Prediction of negative symptoms of schizophrenia from emotion related low-level speech signals. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6024–6028.

FabFilter Software Instruments. (2018). *FabFilter* (Fabfilter pro-q 2.) [Computer software].

Friedman, J., Hastie, T., & Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software*. *33(1)*, 1–22.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

iZotope Inc. (2018). *IZotope RX 6*.

Olsen, L. (2018). *Automatically diagnosing mental disorders from voice: A deep learning approach*.

Olsen, L. (2020). *groupdata2: Creating Groups from Data* (1.3.0) [Computer software]. https://CRAN.R-project.org/package=groupdata2

*Statistics Denmark*. (n.d.). Retrieved 11 November 2020, from https://www.dst.dk/en/

Zivetz, L. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines* (Vol. 1). World Health Organization.

# 8. Appendix

s